图书情报工作
第65卷 第5期 2021年3月

档案学视角下的科学数据管理*
——基于国际组织相关成果的研究

■ 王宁[1,2] 刘越男[1,2,3]

[1] 中国人民大学电子文件管理研究中心 北京 100872 [2] 中国人民大学信息资源管理学院 北京 100872
[3] 中国人民大学数据工程与知识工程教育部重点实验室 北京 100872

摘 要：[目的/意义]在全球 e－science 发展背景下，科学数据管理实践日益呈现出对跨学科思维和方法的渴求，运用档案学领域的相关理论和方法有利于提升科学数据保存和共享重用的质量和效率。[方法/过程]采用文本分析法和综合集成法，对 OCLC、DCC、RDA、ICA 四个国际组织相关文献成果中涉及的档案学理论和方法及相关科学数据管理工作进行了文本编码和归纳分析。[结果/结论]档案学视角下的数字文档连续性保障、背景信息管理、鉴定处置和长期保存对科学数据管理具有支撑作用，建议通过开展跨学科合作对话、建立跨机构连续性管理制度框架、培育具有档案专长的数据馆员等路径提升科学数据管理效能。

关键词：科学数据管理 档案学 国际组织 跨学科 数据馆员
分类号：G203 G275
DOI：10.13266/j.issn.0252－3116.2021.05.009

随着 e-science 的不断发展，科学数据共享和重用成为全球科学界共同的目标。记录科学活动过程的科学数据具有数据和档案的双重属性，从理论上来说，档案学可以为科学数据管理提供理论与方法支撑。在科学数据和相关信息资源融合共享的趋势下，科学数据管理领域和档案领域的权威性国际组织，如联机计算机图书馆中心（Online Computer Library Center，OCLC）、英国数字管护中心（Digital Curation Center，DCC）、研究数据联盟（Research Data Alliance，RDA）和国际档案理事会（International Council on Archives，ICA）等，十分重视档案管理领域的理论与方法在科学数据管理中的应用，通过成立专业兴趣组、开展调查研究、发布指南工具等途径，对档案学理论与方法在科学数据鉴定、全程管理、背景信息管理、长期保存等方面的关键作用开展了系列研究，鼓励引入档案学理论与方法完善科学数据的管理实践，形成了可供借鉴的成果。与此同时，国内外学者也从不同视角提出了档案学理论和方法在科学数据管理中发挥价值的探索性论点，包括档案工作者与科学家合作以了解数据管理和保存需求[1]，档案学原则和技能如来源原则、鉴定和评估、真实性、元数据、风险管理和信任等在科学数据管理中起着至关重要的作用[2]，明确在元数据中捕获体现科学数据质量（准确性、可靠性、真实性等）的内容[3]，促进档案管理员参与科学数据生命周期早期阶段的管理[4]，开展科学数据的价值鉴定等[5]，倡导档案专业人员在科学数据管理中积极发挥作用。

然而，上述研究成果相对零星，缺乏综合集成，且多为西方国家制度和管理环境下的产物。我国档案工作者在科学数据管理工作中的参与度普遍较低，档案学理论与方法在科学数据管理领域应用较少。为进一步增强科学数据管理领域中跨学科方法的融合，笔者认为有必要针对当前科学数据管理面临的挑战，综合科学数据管理领域和档案学领域的相关研究成果，分析档案学对科学数据管理的支撑作用，以期推动图情档跨二级学科的研究，为相关实践提供启发。

1 研究方法与数据来源

本研究主要采用文本分析法和综合集成法。首

* 本文系中国科学院档案馆委托项目“国内外档案工作情况调研”（项目编号：2019K20323）研究成果之一。
作者简介：王宁（ORCID：0000－0002－5070－7624），博士研究生；刘越男（ORCID：0000－0002－5216－2111），教授，博士生导师，通讯作者，E-mail：liuyuenan@ruc.edu.cn。
收稿日期：2020－06－28 修回日期：2020－10－11 本文起止页码：88－97 本文责任编辑：杜杏叶

先,通过网站调研,选择 OCLC、DCC、RDA、ICA 等权威性国际组织发布的相关研究成果为重点分析文本,辅以国内外学者的代表性文献(见表 1),对其中涉及到的档案学理论和方法及其所支撑的科学数据管理工作进行编码(见表 2)。在此基础上,按照档案学理论和方法所解决的问题进行综合集成,归纳可以支撑科学数据管理的档案学视角。

在重点调研的四个国际组织中,OCLC 成立于 1967 年,是联合全球图书馆社区建设的联机计算机图书馆中心,创建了世界上最大的在线公共访问目录 WorldCat[6]。作为全球最大的文献信息服务机构之一,其设计开发的联机计算机系统等产品和服务广泛应用于世界各地的图书馆和科研机构。针对科学数据的管理问题,OCLC 开展了一系列与档案学相关的研究性活动,包括成立专业咨询组、设计面向档案馆和特藏部门的研究学习议程,发布《档案优势:将档案专业知识融合到数字图书馆资料的管理中》研究报告,关注数据用户对数据背景的需求及学术文件管理的发展变化等。

DCC 是国际公认的数字管护专业研究机构,专注于建立数据管理的能力和技能,旨在为存储、管理、保护和共享数字研究数据的机构提供专家建议和实用帮助[7],其设计的数据管护生命周期模型具有广泛的国际影响力。DCC 重视档案鉴定等专业理论在科学数据管理中的重要价值,开发了《研究数据鉴定与挑选指南》《决定数据保存的五个步骤》《在哪里保存研究数据》等指南工具,为科学数据管理提供实操性指导。

RDA 是由欧盟委员会、美国国家科学基金会、美国国家标准与技术研究院以及澳大利亚政府创新部于 2013 年发起的一个社区驱动的国际组织,旨在通过建立社会和技术基础设施,实现全球科学数据开放共享和重用的目标[8]。RDA 成立了档案与文件专业兴趣组(Archives and Records Professionals for Research Data IG,简称 ARPRD)[9],探索以档案、文件管理为代表的信息科学与研究数据管理的交叉领域,倡导将档案专业在元数据、背景信息管理、鉴定和长期保存等方面的技能和优势引入科学数据管理。

ICA 是档案领域最具权威的国际组织,致力于文件档案的有效管理和世界档案遗产的保护利用。其下设的大学与研究机构档案处科学与研究数据委员会专门从事高校科学数据和文件管理的研究。该委员会发布了《科学文件和数据管理与保存指南》,提出了基于研究流程的科学数据识别与管理方案、科学数据长期保存的鉴定标准和管护策略。

表 1　调研文本对象基本情况

编号	文本名称	文本类型	来源
L1	《档案优势:将档案专业知识融合到数字图书馆资料的管理中》[10]	研究报告	OCLC
L2	《社会科学家对数据重用的满意度》[11]	论文	OCLC
L3	《数据重用用户视角的“背景”》[12]	论文	OCLC
L4	《不断发展的学术文件》[13]	研究报告	OCLC
L5	《研究数据鉴定与挑选指南》[14]	指南工具	DCC
L6	《决定数据保存的五个步骤》[15]	指南工具	DCC
L7	《数据管护生命周期模型》[16]	指南工具	DCC
L8	《在哪里保存研究数据》[17]	指南工具	DCC
L9	《RDA 第 11 次全体会议之联合会议:档案与文件专业组和图书馆员专业组》[18]	会议记录	RDA
L10	《RDA 第 9 次全体会议:档案与文件专业组》[19]	会议记录	RDA
L11	《科学文件和数据管理与保存指南》[20]	指南工具	ICA
L12	《档案概念在数据密集型环境中的应用:与科学家合作以了解数据管理和保存需求》[1]	期刊论文	*Archival Science*
L13	《科学数据如何增值? 数字管护与人为因素:文献综述》[2]	期刊论文	*Archival Science*
L14	《今天的数据是明天研究的一部分:科学中的档案问题》[3]	期刊论文	*Archivaria*
L15	《将档案实践向前端移动:基于协作田野调查方法对生态传感数据生命周期的研究探索》[4]	期刊论文	*International Journal of Digital Curation*
L16	《科学数据价值鉴定研究进展》[5]	期刊论文	《情报科学》
L17	《文件管理与研究数据:观点回顾》[21]	期刊论文	*Records Management Journal*
L18	《开放研究数据:问题与机遇》[22]	期刊论文	*Records Management Journal*
L19	《开放研究数据,一个档案挑战?》[23]	期刊论文	*Archival Science*
L20	《数字监护研究中档案学理论的应用及启示探析》[24]	期刊论文	《档案学通讯》
L21	《利益相关者视角下档案部门参与科学数据管理的分析》[25]	期刊论文	《档案天地》

此外，笔者在国际档案学领域和数字管护领域的知名期刊 *Archival Science*、*The American Archivist*、*Archivaria*、*Archives and Records*、*Archives and Manuscripts*、*Records Management Journal*、*International Journal of Digital Curation* 的官网以“scientific data/research data + archive/records”为关键词检索，以及在中国知网以“档案+科学数据/研究数据/数据管护/数字监护”为关键词检索，选取了检出文献中涉及运用档案学理论和方法进行科学数据管理研究的代表性论文，作为本文综合集成的文献源补充。

本研究采用开放性编码的方式对文本数据进行了分析和整合。首先，笔者分别对所有文本进行阅读分析，提取资料中的主要理论和方法概念进行编码；进而对比两份编码结果，将提取的概念进一步归类为更高层次的视角概念，并对其分解和描述。最后，将提取的概念互相贯穿和关联，形成编码对照表，如表 2 所示：

表 2　档案学理论与方法、科学数据管理工作对照编码及对应来源

档案学视角总结	档案学理论与方法	科学数据管理工作	对科学数据管理工作的支撑	来源文本对象
数字文档连续性保障(A2)	前端控制 全程管理 实时捕获	数据连续性管理(S2)	前端管控 全流程管理 数据全生命周期管理	L11,L12,L15,L17,L21
背景信息管理(A3)	背景管理 关联管理 来源原则 元数据	数据背景管理(S3)	保存数据背景 过程数据与结果数据的关联 数据可追溯 元数据	L1,L2,L3,L4,L13,L14,L15,L20,L21
鉴定处置(A1)	价值鉴定 技术鉴定 处置 制定保管期限表	数据选择与处置(S1)	数据鉴定 成本效益权衡 处置 确定保管期限	L1,L5,L6,L7,L11,L13,L16,L17,L18,L20,L21
长期保存(A4)	保存 长期保存	数据保存(S4)	数据管护 长期保存	L1,L8,L9,L10,L11,L12,L14,L17,L18,L19

此外，笔者还对中国科学院高能物理研究所、全国地质资料馆、国家生物信息中心等单位进行了实地调研，以了解其科学数据管理现状和档案部门参与的实际情况，从实证的角度考察和夯实本文综合集成结果的客观性。

2　研究发现

2.1　数字文档连续性保障与科学数据管理工作

2.1.1　数字文档连续性保障

数字文档连续性保障可以理解为数字环境下的文档一体化管理理念，即在从文件产生到销毁或作为档案永久保存的整个文件生命周期中采用连续一致的方法，以减少不同生命周期阶段因管理不一致而产生的内部损耗，达到整体效益最佳。孕育于 20 世纪 40 年代的文件生命周期理论初步揭示了不同阶段文件和档案管理活动的关联[26]。随着电子文件的普及，20 世纪 90 年代澳大利亚著名档案学者 F. Upward 和 S. Mckemmish 提出了文件连续体理论，强调文件档案管理活动的整体性和连续性，在全球档案界引起广泛共鸣，促进了数字连续性政策和行动计划的发展，如英国国家档案馆 2007 年启动的数字连续性项目、新西兰国家档案馆 2009 年启动的数字连续性行动计划和澳大利亚国家档案馆 2015 年发布的《2020 数字连续性政策》，强调在信息持续运动中构建信息管理的系统性框架[27]。在我国，档案学者也逐渐意识到，在文件生命周期后端被动等待业务输出不利于档案质量把控和长期维护，提出了前端控制和全程管理的原则[28]，倡导通过法规、标准、系统、技术等多种方式，在文件形成阶段（甚至提前到系统设计阶段）就介入管理，并对文件创建、捕获保存、处置、组织和利用的整个过程进行持续管控，从而持续保障文件真实性、完整性和可用性。

2.1.2　对科学数据管理工作的支撑

在我国科学数据管理实践中，数据管理部门一般是在科研活动取得阶段性或最终成果之后才对其科学数据进行收集保存，提供共享利用，未从源头上建立全流程、连续性的数据全生命周期管理模式。全国地质资料馆等部门虽然对数据汇交提出了明确的质量要求，但在实践中仍然会面对数据格式多样化的问题，给数据整合和长期保存带来巨大挑战。国务院办公厅 2018 年出台的《科学数据管理办法》（以下简称《办法》）[29]主要规定了科学数据的汇交制度和共享利用工作，并未对科学数据生成阶段的数据规范、收集范围和后期的长期保存做出明确要求，未能体现数据全程管理和前端控制的思想。缺乏连续性管理保障的科学数据管理工作存在数据收集不齐全、数据质量不规范、数据关联不完整、数据利用不充分等潜在问题和风险，不利于科学数据资源的有效增值和开发。

建立数字连续性管理思维，实施全流程连续性管理有助于从根本上提高科学数据质量。可借鉴 DCC 的数据管护生命周期模型和 ICA 的全流程数据识别与管理方案所体现的连续管理思维，同时鼓励档案工作者参与到科研活动早期阶段中，对科学数据开展连续

性管理。

DCC 设计的数据管护生命周期模型是体现连续性管理思维的典型代表。根据该模型,理想的数据管护活动应涵盖从初始的概念化设计到数据创建与接收、鉴定与挑选、数据采集、保存与存储、获取利用、转化与迁移、社区观察和参与、数据描述等各个阶段在内的数据全生命周期。其中,概念化设计即设想并规划数据的创建活动,包括设定捕获方法和数据存储范围;创建和接收是指创建元数据,并从数据创建者、其他单位存储库或数据中心接收数据[16]。这两项活动充分体现了对数据管理的前端管控思维,在数据形成阶段就嵌入数据的管理方案和收集范围,明确元数据要求,全面保障整个数据生命周期内管理的连续性。

ICA 以科学活动生命周期为基础,提出了基于研究全流程的数据识别与保存管理方案,也体现了科学数据连续性管理的思维。该方案将一般研究项目的整个过程概括为八个不同阶段的循环过程(见图 1),分别是科学问题的提出、规划、原始数据收集、分析、评估审计、结果报告、财务报告和催生新研究[20]。在整个研究活动过程中,存档被视为其中的一项核心活动,除研究问题提出之外的每个阶段都要对产生的数据进行收集和保存,进而形成完整的科学数据流。该方案倡导以归档为核心开展数据管理活动,体现了实时捕获、同步管控、集成管理的理念,有助于确保各类科学数据收集、保存和管理的连续性。

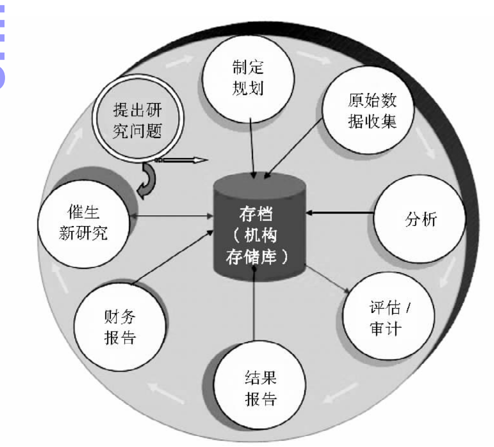


图 1　研究活动阶段分布与数据流向示意[20]

此外,e-science 研究活动的成功不仅取决于科学家和技术人员之间的有效合作,也取决于档案管理员的积极参与,应鼓励档案工作者更好地理解科学活动过程,并使其能够参与到数据生命周期的早期阶段,从前端改善科学数据的管理质量。[4]

2.2　背景信息管理与科学数据管理工作

2.2.1　新来源观与背景信息管理

来源原则(Principle of Provenance)是世界公认的档案整理理论,也是档案学的支柱性理论之一。来源原则强调档案整理要尊重来源,尊重全宗的完整性以及尊重全宗内的原始有机联系。[26] 按照来源而非主题组织档案信息成为档案领域的独特方法。在电子文件时代,来源原则受到新技术环境的冲击,进而导致来源原则“重新发现”,并诞生了“新来源观”。新来源观视角下,学者们对“来源”的概念进行了重新阐释,突破“文件形成者来源”或“机构来源”的固有认识,将文件的形成背景(Context),即文件是由谁、在什么职能活动中、为了何种目的、采用怎样的结构形式生成等方面的综合背景信息也视为来源信息[26]。背景信息不仅是档案组织的依据,也是档案鉴定的重要参考,即档案鉴定不仅要判断单份档案的价值,还要根据档案之间的背景关联来判断同一个业务活动中产生的一整套档案的价值。从文件构成来看,背景和内容、结构是文件的三要素。档案界一致认为背景是文件档案之所以成为业务活动凭证的关键,是维护电子文件真实性、完整性和可理解性的重点,特别强调关注形成文件的职能、计划、活动、业务等表现“宏观联系”的背景信息。因此,背景是档案学理论的核心概念,背景信息管理是档案管理的核心技能之一,档案工作者在捕获数字文件本身的同时,需要同时捕获文件创建过程、利用权限、保管情况和预期用途等背景信息。

2.2.2　对科学数据管理工作的支撑

在科学数据整理方面,人们会认为按照学科或主题分类是常见做法。然而,笔者调研 3 个案例中,有 2 个遵循了来源原则。其中地质资料馆的做法是将一个调查活动中形成的所有资料整理成卷,该馆同时具备科学数据中心和专业档案馆的性质,其做法源自对于科技档案“成套性”整理的要求。国家生物信息中心按照“项目(project)－样本(Sample)－实验(Experiment)－测序(Run)”的结构组织生物科学数据,在整理逻辑上与按照文件业务来源的背景进行整理的档案学方法高度一致。虽然该单位的科学家和数据管理者并不了解档案学理论和方法,但其在实践中摸索的信息组织方法恰恰体现了背景关联旺盛的生命力。

在数据溯源方面,随着大数据技术的不断发展,通过整合不同领域、不同来源和不同类型的科学数据,进行综合分析来解决科学问题的趋势越来越显著,这些数据的真实性和可信性直接影响到分析结果的准确

性,因此科学数据信息的背景可追溯性变得日益重要。OCLC 通过对数据重用满意度和数据用户所关注的数据背景的研究发现,从科学数据重用需求的视角来看,保存数据产生的背景信息和保存数据内容同样重要,且数据的几个质量属性——完整性、可访问性、易操作性和可信度,与数据重用满意度有着明显的正相关关系[11-12]。因此,科学数据的保存不仅要保存结果数据和过程数据本身,还要保存数据软件信息、数据提供者、研究项目信息、处理活动、共享利用等背景信息,否则未来研究者可能找不到完整的资料支撑新的研究,反映相关研究历史的记录也可能残缺[13]。OCLC 指出档案工作者在背景信息管理方面具有优势,从初步调查到资料处理和元数据的创建,对文件生命周期每个阶段的背景理解都至关重要,这在研究领域同样重要[10]。在科研成果发现、发表之前,如果缺失了档案工作者的参与,可能导致丢失有关数据来源、背景和项目的重要信息[4]。

数字环境下,相关背景信息通常通过元数据来体现,元数据是数据规范化管理的基础,也是数据管理计划的重要组成部分。在数字档案资源管理中,只要始终掌握其形成、管理与利用的元数据,并与档案资源内容相互关联,就可以掌握其生成与运转的来龙去脉,从而有效维护档案资源的历史联系[24]。元数据也是科学数据管理的基本工具。然而,科学数据生成者——科研人员往往缺乏元数据"驱动着数据管理生命周期中所有步骤"的认识,描述背景的元数据供给不足,需要数据管护人员补充改善[2]。此外,背景元数据往往随着数据管理过程不断产生,很难事后补录。ICA 提出在项目规划阶段就应将存档元数据声明等设计在内,在数据保存和管理过程中尽可能捕获工作流程数据,保障科学数据及其背景信息的及时和完整保存,避免有价值的背景信息的缺失[20]。

2.3 档案鉴定处置与科学数据管理工作

2.3.1 档案鉴定处置

"鉴定是最崇高的职能,是当代档案实践的核心"[30]。在档案学领域,鉴定又称价值鉴定,是指判断原始业务信息(即文件)在业务结束后是否仍然具有保存价值的工作。鉴定关系到档案管理对象的选择,是最核心和关键的档案管理活动,包括对文件信息的价值进行评估,判断其在业务、制度、法律、财政、历史等方面的价值及其对未来的潜在使用价值,从而判定其是否属于归档范围并确定保管期限的过程。档案学在长期发展过程中,形成了具有很强理论根基的价值鉴定理论,先后建立了高龄档案鉴定论、职能鉴定法、直接鉴定法、利用需求预测法等多种鉴定方法。数字时代职能鉴定法得到全世界档案理论研究领域和实践领域的广泛认同,并在中、美、澳、加等多个国家的档案鉴定政策中得以体现。档案部门会根据鉴定方法的应用结果,精心设计保管期限表来支持档案保管期限的划分以及档案的处置工作,包括将具有长期保存价值的档案移交到档案馆进行长期或永久保存,对保管到期的档案进行销毁等[22]。

2.3.2 对科学数据管理工作的支撑

随着科学数据在各类科研活动中的急剧产生,海量科学数据存储的成本和效益问题浮出水面:一方面,随着数字内容的不断扩展,尽管数据存储载体的成本有所下降,但是数据备份、元数据维护、格式管理、质量检测等数据维护的成本成倍地增加。只有当科学数据自身所具有的价值大于其管理成本时,才有必要加以续存,然而并非所有的科学数据都具备这样的潜在价值;另一方面,保存所有数据会给数据检索和利用带来巨大的挑战,保存的内容越多,检索的信噪比越高,数据用户精确获取目标数据的效率就越低。因此,开展科学数据的鉴定非常必要,笔者所调研的中国科学院高能物理研究所的数据管理人员就提到,在数据管理实践中,虽然大量科学数据都可能具有保存价值,但是出于经费原因,只能选择其中具有重要价值的数据进行保存,且保存的时间长短也将取决于经费的支持情况。DCC 也指出"科学数据存储的规模非常大,而且需要保存足够的元数据以确保数据随时间推移可追溯、可理解和可用。考虑到长期保存和管理数据需要承担的未来费用,数据创建者和管理者都无法逃避做出鉴定决策"[14]。然而,实践中最重视的仍是近期的科学数据汇交和共享利用,对相对长远的鉴定处置等问题仍缺乏有效经验,鉴定的责任主体亦不明确[5]。档案鉴定处置方法有助于支持科学数据管理者有效选择数据,判定其保存价值和保管期限。

国际组织相关成果对档案鉴定处置方法在科学数据的挑选和成本效益权衡等方面的价值达成了共识,均将可用性、重用价值、数据质量等作为科学数据鉴定的重要参考标准,并对数据鉴定主体和技术操作思路进行了探讨,形成了具有参考价值的成果,代表性观点包括 DCC 的综合价值评估法和五步骤实施策略,ICA 的鉴定三标准和文件数据保管处置方案,OCLC 的分阶段鉴定思路等。

DCC 明确将"鉴定和挑选"作为管护生命周期的

八项活动之一，要求数据管理者“鉴定并挑选数据以进行长期管护和保存”[16]，建议“研究机构的数据馆员和档案管理员主要负责制定挑选和鉴定政策，参考数据生成者、数据重用者、研究社区等利益相关方的意见制定政策”[14]。其中，研究机构制定的鉴定政策需要规定评估数据集价值的七个标准，分别为：与科研机构使命的相关性，数据的科学、文化或历史价值，数据独特性，数据质量，数据的不可复制性，经济成本，著录的完整性，并判断数据的保管期限和销毁时间[14]。此外，DCC 还提出通过考虑潜在的数据重用需求、检查各类数据指标（确保满足法律和政策要求）、确定具有长期保存价值的数据、权衡经济成本、制定保存或处置行动等五个具体步骤开展数据的鉴定工作[15]。

ICA 倡导的数据鉴定要遵循三个基本标准，保证数据的可信度、有效性和质量：“①数据要具备用于验证结果的必要性；②要能确保数据保存后访问获取的可行性；③数据要有重用和创建新研究的可能性”[20]。在鉴定的基础上，对应研究活动各阶段科学数据的产生情况，ICA 为除研究问题提出之外的每个阶段都制定了一份文件与数据保管与处置方案（示例见表 3），详细规定了要收集的文件和数据类型、载体格式、保管和处置要求及利用限制等主要内容，提供了直接可参考的科学数据归档范围及保管期限规范[20]。

OCLC 提出鉴定可以分一个或多个阶段进行。鉴于电子档案鉴定在对档案内容有用性的价值鉴定之外，增加了对其是否处于可用状态的技术鉴定，OCLC 建议在存储机构与移交者进行交接前或在材料被收集保存之后进行价值鉴定，在收集之前对包含原始数字信息的材料进行技术鉴定，即使用适当的数字工具进行检查，审查内容是否损坏和篡改等[10]。

表 3　数据收集阶段文件与数据保管与处置方案[20]

文件/数据	格式	行动	备注
观测数据	数字；纸质	①根据保存标准长期保存；②如果需要鉴定，处置方式必须记录在案；③根据约束规则进行公共获取或限制访问	①只要用于验证研究结果，一般至少保存 10 年，有关毒品研究的内容保存 15 年；②数据获取应遵循国家规定或机构、学科规则
实验数据	数字；纸质	同上	同上
有关实验和观测的协议书	数字；纸质	同上	同上
实验室笔记	数字；纸质	同上	同上
期刊	数字；纸质	同上	根据国家规定实施
仿真文件（如数据模型）	数字；纸质	同上	①只要用于验证研究结果，一般至少保存 10 年；②数据获取应遵循国家规则或机构、学科规定
序列数据	数字；纸质	同上	同上
……			

2.4　数字档案长期保存与科学数据管理工作

2.4.1　数字档案长期保存

作为信息管理领域的重要任务，数字信息的长期保存已经引起了图书馆学、档案学和数据科学等多学科领域的共同关注和实践中多部门的共同推进。其中，档案部门基于长期管护社会记忆资产的职责所在，致力于保障文件档案信息的长久可用和长久可信。经过多年探索，国际档案领域在数字档案信息的长期保存实践方面已经积累了丰富的经验，并形成了一些独特的技术路线，如英国国家档案馆开发的数字格式登记系统 PRONOM 项目[31]、瑞士联邦档案馆开发的基于 XML 进行长期保存关系数据库的 SIARD 方案[32]、澳大利亚维多利亚州采用的元数据封装方案（VEO）[33]等，均产生了广泛的国际影响力。国际上已经普遍认同将数字档案馆认定为数字存储库的重要类型，为数据集提供存储和访问平台，开展标准化的数据质量控制和完整的生命周期管理[25]。

2.4.2　对科学数据管理工作的支撑

科学家越来越认识到，他们缺乏满足数据保存所需的技能和专业知识，正在寻求“数据档案管理员”的帮助，因为对档案资源的收集、组织和长期保存是档案工作者的专业使命[1]。国外许多科研资助机构和科研管理机构都将“数据归档和长期保存”列为数据管理计划的重要组成部分[34]，而我国的《办法》仅对科学数据保存提出原则性要求[29]。根据笔者调研，实际单位目前对长期保存虽有思考，但主要采用备份等基本策略，缺乏迁移、仿真、保存元数据等核心策略的应用。2019 数字资源长期保存全国学术研讨会上，有专家指出档案工作者在档案管理实践中就长期保存形成的方法经验和实践成果可以为科学数据的长期保存提供一定的参考。

DCC 提供了数据存储的可选方案指南，指出有上

百种可以进行数据存储的存储库，各有不同的优缺点，为开放获取选择存储库和为长期保存选取存储库所考虑的因素是不同的。就开放获取和共享数据而言，学科特定的数据存储库、科学数据中心、通用数据存储库、机构数据存储库、期刊补充资料服务、网站等可以作为存储数据的选项；而就数据的长期保存而言，长期保存的成本、安全性和可用性是数据保存的重点因素，建议通过综合考量，选择机构数据档案库、安全中心、云存储、数据存档第三方服务等方案进行长期保存[17]。

ICA 指南从档案学专业角度出发，提出了科学数据和文件长期保存与管护方面的基本标准和策略，强调科学数据、文件档案管理要和科研活动的流程集成，防止造成后期无法弥补或者成本过高的损失和风险。此外，应通过迁移、仿真或以原生格式保存等不同策略实施长期保存[20]。RDA 的 APARD 小组也十分关注科学数据的长期保存问题，其第 9 次全体会议的主题即"聚焦数字保存"[19]，第 11 次全体会议上提出起草有关数字保存的简要指南，收集 ARPRD 成员和其他小组有关科学数据相较于其他数字资产在长期保存中所面临的特殊挑战的观点，并依据美国国家数字管理联盟（National Digital Stewardship Alliance，简称 NDSA）的"数字保存级别"文档（The NDSA Levels of Digital Preservation：An Explanation and Uses），讨论潜在的更新或修订思路[18]。

在国外科学数据管理实践中，已经有研究机构与档案馆合作或建设数字档案馆开展数据长期保存的案例，也为科学数据长期保存提供了档案机构参与的实践经验。如伊利诺伊大学香槟分校图书馆研究数据服务中心（Research Data Service，简称 RDS）与大学档案馆合作，承诺在 RDS 出版数据后至少 5 年内保存并促进对数据集的访问，在 RDS 接收研究数据五年之后，基于档案学的鉴定理论，再决定继续保留、增加资源抑或销毁[35]。美国国家科学基金会资助的大气研究中心（National Center for Atmospheric Research，简称 NCAR）建立了研究数据档案馆，用于支持长期保存具有不可替代性的科学数据，以及超过 40 年的异构存档数据，通过持续更新 IT 技术以增强数据发现和访问能力，为 NCAR 研究人员提供数据管理支持[36]。

3　研究讨论

基于国际组织和学界运用档案学理论与方法参与科学数据管理工作的研究，结合我国当前科学数据管理面临的档案管理视角缺失的现实问题，建议通过开展跨学科协作对话、建立跨机构连续性制度框架、培养具备档案专长的数据馆员等路径，促进运用档案学知识技能提升科技信息资源的整体管理效能。

3.1　开展科学数据管理的跨学科协作对话

国际组织已经在科学数据管理的跨学科对话上取得了相应进展，如 OCLC 专门设计了在研究图书馆系统内面向档案馆和特藏部门的研究学习议程（Research and Learning Agenda for Archives，Special，and Distinctive Collections in Research Libraries），成立了由档案馆和特藏部门负责人组成的咨询组，了解在整个科研管理生态系统内跨部门、跨专业领域的不同管理问题和知识需求，促进宣传和开发档案馆和特藏部的资源，由咨询组的专家成员在整个研究过程中定期提供咨询和意见[37]。RDA 也提出档案工作者、文件管理专业人员和图书馆员长期以来一直共同致力于获取、鉴定、编目、管理、保存和提供获取数字和模拟的研究材料，这些专业人员都拥有可以为最佳实践的发展做出巨大贡献的技能和专业知识，联手协作将更有助于良好的科学数据管理和共享目标的实现[38]。在 RDA 联盟实践中，ARPRD 就与图书馆员兴趣组（Libraries for Research Data IG）合作，在第 11 次论坛上举行的联合会议上，共同探讨两个小组可以合作的项目及主题，包括研究数据的鉴定、数字保存、元数据等，希望通过小组合作的方式推动对研究数据管理相关领域的发展。两个专业组还致力于合作开发科学数据管理基础设施和最佳实践，以确保在五年、二十年、五十年、一百年或更长时间内可以访问和使用数据集[39]。

虽然我国从 20 世纪 80 年代起就倡导图书、情报和档案的一体化管理和发展，但是在学科研究和发展实践中，壁垒仍然明显存在。除了缺乏跨领域的机构合作之外，我国当前尚未建立融合图书馆学、情报学、档案学、数据科学等信息学科的综合研究协会组织，未来可以考虑加强该综合学科领域的合作建设。同时以国际组织的合作研究和兴趣小组机制为启发，建议我国的图书馆学会、档案学会和科学技术情报协会等充分利用现有的合作平台，或在国际科技数据委员会中国委员会中设立相关研究兴趣小组，就多学科共同关注的数据长期保存、数据鉴定、元数据、数据存储库等问题加强探索创新，增进对其他学科特长的了解，合作促进开发和完善相关科学数据管理基础设施，以服务于科学数据的全生命周期管理。

3.2 建立跨机构的科学数据连续性管理制度框架

在档案学领域,连续性管理思维除了强调信息对象的全流程连续之外,还强调文件档案管理从文件形成单位到档案馆的管理连续性,形成一个完整的跨机构制度框架,也可以为科学数据的管理提供全流程连续性管理视角,促进科学数据管理的多主体合作,加强管理的连贯性。

从资源视角来看,需要促进科学数据和科研档案的集成管理和服务。科学数据和科研档案在对象上存在交叉,但我国的科学数据和科研档案的管理长期处于"割裂"状态,既没有将档案管理环节纳入科学数据的管理过程,也没有将科研档案的共享利用纳入科学数据共享利用中。科研院所的科学数据管理和档案管理一般由不同职能管理部门承担,二者存在着显著的分工差异和不同的业务侧重,尚未形成从科学数据生成到归档保存的完整链条。随着近年来国际科学界日益重视科研成果、科学过程数据和科研管理档案的集成共享利用,客观上也对存储在档案馆的科研档案和存储在科学数据中心的科学数据提出了整合服务的需求。由此,有必要建立科学数据管理机构和档案部门的协同工作制度,档案部门与科学数据中心就数据汇总格式、数据提交规范、数据管理方案和长期保存计划等进行协商及研究[25],推动两项工作的融合发展;同时不断推进科研档案的数据化和资源整合服务,打破"信息孤岛",增强科技信息资源管理和服务的整体水平。目前我国仅有全国地质资料馆等少量具有科技信息资源一体化管理职能的机构,在开展科学数据管理的同时承担着档案馆的功能,应大力支持和推广此类机构的协同发展模式,促进科学数据和科研档案的协同管理。

从管理视角来看,科学数据的管理并非仅仅是科研机构的任务,其生成到保存可能需要跨机构开展,同样需要构建跨机构的管理框架。国外科学数据管理领域已经将档案馆视为一种重要的数据存储库类型。随着我国数字档案馆的不断建设,档案机构也具备了一定的长期保存数字信息资源的能力,建立了较为成熟的长期保存技术策略,可以作为科学数据存储库的分担者,与科学数据中心等共同承担科学数据的保存、管理工作,尤其是具有重要的社会、历史、文化价值的科学数据,可以选择移交到档案馆进行保存。在此基础上,档案工作者将有机会成为存储库管理者、数据馆员或数据科学家,从而发挥自身的技能和专长参与到科学数据管理工作中[22]。

3.3 培养具备档案专长的数据馆员

随着 e-science 和 open science 的快速发展和科学数据管理需求的增加,国内外科研机构、科研资助机构、学术图书馆与信息中心等科学数据管理机构出现了一个新的岗位类型,即实施科学数据管理、开展数据监管、服务数据开放利用的数据馆员。虽然作为一个新兴的职业类型,学界尚未给出一致的定义,但是在科学数据管理实践领域已经呈现出高需求,国际社会科学信息服务与技术协会 2017 年收录的 64 条招聘信息中,就有 41 个岗位与数据馆员相关[40]。顾立平等提出,"开放科学环境下的数据馆员,应是运用图书馆工作原理、具备科学数据管理知识技能,了解开放科学运行机制和特定研究领域知识背景的数据管理从业人员"[40]。这一定义首先强调了科学数据管理中图书馆工作原理的重要性,但没有明确档案学理论与方法的必要性。笔者认为,档案学知识技能应自动包含在"科学数据管理知识技能"之内,即充分理解数字文档连续性管理思维、了解背景信息管理需求、熟悉数据鉴定原则和掌握数据长期保存技能。

因此,有必要加强培养具备档案专长的数据馆员。首先,可以针对科研机构、科研资助机构、学术图书馆和信息中心等既有的数据馆员,组织开展档案学相关知识技能的培训和指导,打开其运用档案思维开展科学数据管理的视野。如美国国家档案馆和文件管理署开设了数据管护与培训项目,以及英国数据档案馆专门为社会科学领域的学者提供培训服务等均是档案界近年来积极开展数据管理较为成功的案例[25]。此外,可以在高等院校、研究院所开设的图书馆学、档案学、数据科学、信息资源管理等学科学位教育中,为有志于从事科学数据管理工作的学生提供数据管护及档案学相关的必修、选修课程,完善其知识结构,培育数据管理综合素养。如加州大学洛杉矶分校、印第安纳大学、西蒙斯学院、马里兰大学、密歇根大学等多所国际著名高校的信息学院在图书情报学或档案学硕士培养方案中设立了数字文件与信息管理、数字保存、数据管护、元数据、可信数字仓储等相关课程,注重全面培养学生的数字信息管护技能。最后,相关机构聘用数据馆员时,应将具备档案学专业教育背景的毕业生或具有档案从业经验的人员纳入招聘范围,以丰富科学数据管理的人才结构。只有将档案学知识技能纳入到数据馆员培养和聘用的需求框架之内,才能真正发挥档案专长,更好地服务和完善科学数据管理工作。

4 结语

档案学理论与方法在科学数据管理领域具有独特的专业价值，且日益受到科学数据管理研究领域的重视。鉴于国内档案机构和档案从业人员在科学数据管理中参与度较低、科学数据管理中档案视角相对缺失的现象，有必要借鉴国际组织的相关研究成果，推动加强科学数据管理的跨学科、跨领域、跨机构的协作交流，充分挖掘档案学的优势，培养具备档案专长的数据馆员，使其参与到制定科学数据鉴定方案、长期保存规范、元数据方案和连续性管理制度的实践中，进而推动科学数据管理和共享服务的提质增效。

参考文献：

[1] DHARMA A, ANN Z, MORGAN D, et al. The application of archival concepts to a data-intensive environment: working with scientists to understand data management and preservation needs[J]. Archival science,2011,11(3/4): 329-348.
[2] ALEX H. How has your science data grown? Digital curation and the human factor: a critical literature review[J]. Archival science, 2015,15(2):101-139.
[3] TRACEY P, BARBARA L, FRASER T, et al. Today's data are part of tomorrow's research: archival issues in the sciences[J]. Archivaria,2007,64(Fall):123-179.
[4] JULLIAN C, CHRISTINE L, MATTHEW S. Moving archival practices upstream: an exploration of the life cycle of ecological sensing data in collaborative field research[J]. The international journal of digital curation,2008,1(3):114-126.
[5] 邓君，宋文凤. 科学数据价值鉴定研究进展[J]. 情报科学，2012(6):943-946,958.
[6] WorldCat. org: The world's largest library catalog[EB/OL]. [2020-06-24]. https://www. worldcat. org.
[7] About DDC[EB/OL]. [2020-05-21]. https://www. dcc. ac. uk/about.
[8] About RDA[EB-OL]. [2020-05-22]. https://www. rd-alliance. org/about-rda.
[9] IG Archives and records professionals for research data[EB/OL]. [2020-05-25]. https://www. rd-alliance. org/ig-archives-and-records-professionals-research-data. html.
[10] The archival advantage: integrating archival expertise into management of born-digital library materials[EB/OL]. [2020-06-18]. https://www. oclc. org/research/publications/2015/oclcresearch-archival-advantage-2015. html.
[11] Social scientists' satisfaction with data reuse[EB/OL]. [2020-06-15]. https://www. oclc. org/research/publications/2015/social-scientists-satisfaction-with-data-reuse. html.
[12] Context from the data reuser's point of view[EB/OL]. [2020-06-16]. https://www. oclc. org/research/publications/2019/context-from-data-reuser-point-of-view. html.
[13] The evolving scholarly record[EB/OL]. [2020-06-16]. https://www. oclc. org/research/publications/2014/oclcresearch-evolving-scholarly-record-2014-overview. html.
[14] How to appraise and select research data for curation[EB/OL]. [2020-06-12]. https://www. dcc. ac. uk/guidance/how-guides/appraise-select-data.
[15] Five steps to decide what data to keep[EB/OL]. [2020-06-20]. https://www. dcc. ac. uk/guidance/how-guides/five-steps-decide-what-data-keep.
[16] Curation lifecycle model[EB/OL]. [2020-06-17]. https://www. dcc. ac. uk/guidance/curation-lifecycle-model.
[17] Where to keep research data[EB/OL]. [2020-06-17]. https://www. dcc. ac. uk/guidance/how-guides/where-keep-research-data.
[18] Joint meeting: IG Archives and Records Professionals for Research Data, and IG Libraries for Research Data Interest Groups Plenary11[EB/OL]. [2020-06-18]. https://www. rd-alliance. org/archives-and-records-professionals-research-data-ig-rda-plenary-11-berlin.
[19] RDA P9 Archives & Records IG: notes & slides[EB/OL]. [2020-05-30]. https://www. rd-alliance. org/group/archives-and-records-professionals-research-data-ig/post/rda-p9-archives-records-ig-notes.
[20] New handbook: management and preservation of scientific records and data[EB/OL]. [2020-05-22]. https://www. ica. org/en/new-handbook-management-and-preservation-scientific-records-and-data.
[21] GRANT R. Recordkeeping and research data management: a review of perspectives[J]. Records Management Journal, 2017, 27(2):159-174.
[22] CHILD S, MCLEOD J, LOMAS E, et al. Opening research data: issues and opportunities[J]. Records Management Journal, 2014, 24(2):142-162.
[23] BORGERUD C, BORGLUND E. Open research data, an archival challenge? [J]. Archival Science,2020:1-24.
[24] 毛天宇. 数字监护研究中档案学理论的应用及启示探析[J]. 档案学通讯，2016(1):34-38.
[25] 闫鹏. 利益相关者视角下档案部门参与科学数据管理的分析[J]. 档案天地，2019(3):23.
[26] 冯惠玲. 档案学概论[M]. 北京：中国人民大学出版社，2006:259.
[27] 周文泓，张宁. 全球数字连续性的行动全景与启示——基于英国、新西兰、澳大利亚与美国国家政策的探讨[J]. 情报理论与实践，2017,40(3):138-142,137.
[28] 李铭. 档案化管理前端控制和全程管理的核心[J]. 浙江档案，2005(11):7-8.
[29] 国务院办公厅关于印发科学数据管理办法的通知[EB/OL]. [2020-06-12]. http://www. gov. cn/zhengce/content/2018-04/02/content_5279272. htm.

[30] CAROL C. Archival appraisal: a status report[J]. Archivaria, 2005,59:83 – 107.
[31] 张宁,杨敬敬. 国外典型数字格式登记系统比较研究——以 PRONOM、GDFR 与 UDFR 为例[J]. 北京档案,2015(9):17 – 20.
[32] 钱毅,刘力超. 数据库电子文件归档与长期保存技术路径研究[J]. 档案学研究,2017(4):67 – 72.
[33] 刘越男. 对电子文件元数据封装策略的再思考——由 VERS 标准的变化引起的研究[J]. 档案学研究,2019(4):116 – 123.
[34] 刘峰,张晓林. 数据管理计划构成规范及其可操作数据监护模型研究[J]. 现代图书情报技术,2016(1):11 – 16.
[35] Processing digital research data[EB/OL]. [2020 – 11 – 19]. https://saaers. wordpress. com/2016/05/11/processing-digital-research-data/.
[36] Research data archive, National Center for Atmospheric[EB/OL]. [2020 – 06 – 15]. https://rda. ucar. edu/#! about
[37] Research and learning agenda for archives, special, and distinctive collections in research[EB/OL]. [2020 – 06 – 15]. https://www. oclc. org/research/publications/2017/oclcresearch-research-and-learning-agenda. html.
[38] RDA and librarianship, archival science and information science | RDA [EB/OL]. [2020 – 06 – 22]. https://www. rd-alliance. org/rda-disciplines/rda-and-librarianship-archival-science-and-information-science.
[39] 23 Things: libraries for research[EB/OL]. [2020 – 06 – 22]. https://rd-alliance. org/group/libraries-research-data-ig/outcomes/23-things-libraries-research-data-supporting-output.
[40] 顾立平,张潇月. 开放科学环境下数据馆员的实践探析[J]. 图书情报知识,2020(2):60 – 74,112.

作者贡献说明：
王宁：提出研究问题、制定研究框架、论文撰写；
刘越男：制定研究框架、修正研究框架、论文撰写及修改。

Scientific Data Management from the Perspective of Archives: A Study Based on Relevant Achievements of International Organizations

Wang Ning[1,2] Liu Yuenan[1,2,3]
[1] Electronic Records Management Research Center, Renmin University of China, Beijing 100872
[2] School of Information Resource Management, Renmin University of China, Beijing 100872
[3] Key Laboratory of Data Engineering and Knowledge Engineering, Renmin University of China, Beijing 100872

Abstract: [**Purpose/significance**] In the context of global e-science development, scientific data management practices have increasingly shown a desire for interdisciplinary thinking and methods. The use of relevant theories and methods in the field of archives can help improve the quality and efficiency of scientific data preservation, sharing, and reuse. [**Method/process**] By use of text coding analysis and comprehensive integration method, the archival methods and the involved scientific data management work were extracted and inducted from the research achievements of four international organizations including OCLC, DCC, RDA and ICA, as well as other related literature. [**Result/conclusion**] It is found that the methods of archival science include appraising and disposal, digital continuity, context management, long-term preservation are necessary to carry out scientific data management. It is recommended to improve the effectiveness of scientific data management by conducting interdisciplinary cooperation dialogues, establishing a cross-agency continuity management regulation framework, and cultivating data librarians with archival expertise.

Keywords: scientific data management archival science international organizations interdisciplinary data librarian